
Partiel MAO Probabilités et Statistiques (18 mai 2021, 9h30-11h30)

Lorsque l'énoncé vous demande de **coder**, vous avez le choix entre :

1. Écrire du code Python. Dans ce cas on pourra supposer qu'on a appelé les modules habituels :

```
import numpy as np
import scipy as sp
import matplotlib.pyplot as plt
import numpy.random as rnd
import scipy.stats as sts
```

Vous ne perdrez pas de point si vous vous trompez dans les noms des fonctions et leurs options.

2. Écrire du pseudo-code, par exemple :

```

FUNCTION comptage(p)
  n = 0
  POUR k de 1 à 10
    u = Uniforme([0, pi])
    SI u < p
      n = n+1
  RENVoyer n
```

Exercice 1.— Simulation de lois réelles

La loi de l'arcsinus sur $] -1, 1[$ est la loi de densité

$$f(x) = \frac{1}{\pi\sqrt{1-x^2}} \mathbb{1}_{]-1,1[}(x).$$

La loi du semi-cercle sur $[-2, 2]$ est la loi de densité

$$g(x) = \frac{1}{2\pi} \sqrt{4-x^2} \mathbb{1}_{[-2,2]}(x).$$

On souhaite simuler ces deux lois.

1. Expliquer comment simuler la loi de l'arcsinus par la méthode d'inversion. On rappelle qu'une primitive de la fonction $\frac{1}{\sqrt{1-x^2}}$ est la fonction $\arcsin(x)$.

Coder une fonction qui renvoie une réalisation de la loi de l'arcsinus.

Correction — On calcule la fonction de répartition, on trouve $F(x) = \left(\frac{1}{2} + \frac{1}{\pi} \arcsin x\right) \mathbb{1}_{]-1,1[}(x) + \mathbb{1}_{[1,\infty[}(x)$. Elle est inversible sur $] -1, 1[$, d'inverse $F^{-1}(y) = \sin\left(\pi\left(y - \frac{1}{2}\right)\right)$ où $y \in]0, 1[$. Par la méthode d'inversion, la loi de l'arcsinus sur $] -1, 1[$ a donc même loi que $\sin(\pi(U - \frac{1}{2}))$ où $U \sim \mathcal{U}([0, 1])$. Code possible :

```
def var_arcsin():
    return np.sin( np.pi * (rnd.random() - 0.5) )
```

2. On suppose qu'on a utilisé l'algorithme de la question 1 pour simuler N réalisations indépendantes, (X_1, \dots, X_N) . Proposer deux manières de visualiser le fait que cet échantillon suit bien la loi demandée (on ne demande pas de les coder).

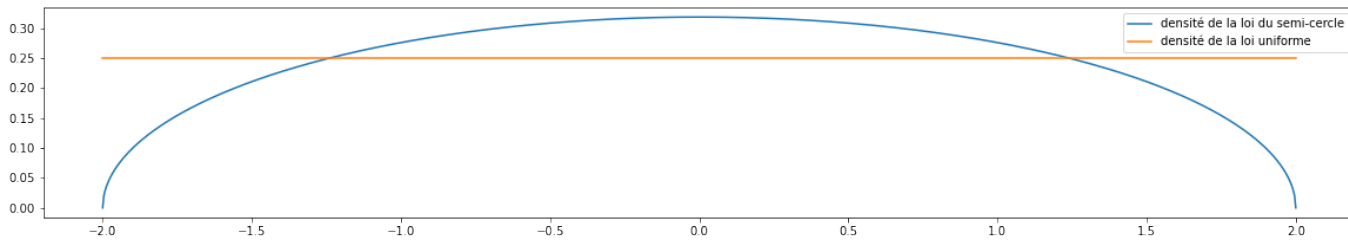
Correction — On peut tracer la fonction de répartition empirique des (X_i) et la superposer à la fonction F trouvée précédemment. Ou bien, on peut tracer un histogramme des (X_i) et le superposer à la densité f .

3. Tracer à la main le graphe de la fonction g , ainsi que la densité de la loi uniforme sur $[-2, 2]$. Supposons qu'on sait simuler une variable uniforme sur $[0, 1]$, expliquer comment simuler la loi du semi-cercle par la méthode de rejet.

Coder une fonction qui renvoie une réalisation de la loi du semi-cercle.

Correction — Pour le tracé, voir la figure suivante. On appelle $h(x) = \frac{1}{4}\mathbb{1}_{[-2,2]}(x)$ la densité de la loi uniforme sur $[-2, 2]$, alors on remarque que $g \leq \frac{4}{\pi}h$. On peut donc appliquer la méthode de rejet : soient (X_n) des variables iid de loi uniforme sur $[-2, 2]$ (qu'on peut obtenir comme $X_n = 4U_n - 2$ où U_n sont iid $\mathcal{U}([0, 1])$), et (Y_n) des variables iid uniformes sur $[0, 1]$, alors les $(X_n, \frac{4}{\pi}Y_n)$ sont des points iid uniformes sous la courbe de $\frac{4}{\pi}h$, et il suffit de renvoyer l'abscisse du premier de ces points qui est sous la courbe de g . Pour tester cela, on teste si $\frac{4}{\pi}Y_n \leq g(X_n)$; on peut raffiner un peu en disant que c'est équivalent à $(8Y_n)^2 + X_n^2 \leq 4$, ce qui est un peu plus rapide à calculer pour l'ordinateur (mais on ne demandait pas d'aller jusque là). Code possible :

```
def var_sc():
    while True:
        xn = 4*rnd.random() - 2
        yn = rnd.random()
        if (8*yn)**2 + xn**2 <= 4:
            return xn
```



4. Quelle est la loi du nombre d'essais dans l'algorithme de la question 3 ?

Correction — Géométrie de paramètre $\frac{\pi}{4}$.

Exercice 2. — Une suite autorégressive

Soit a, b deux nombres réels, et soit $(Z_n)_{n \geq 1}$ une suite i.i.d. de variables aléatoires de loi normale centrée réduite $\mathcal{N}(0, 1)$. On considère la suite de variables aléatoires $(X_n)_{n \geq 0}$ définie par

$$X_0 = 0, \quad \text{et} \quad \forall n \in \mathbb{N}, \quad X_{n+1} = aX_n + b + Z_{n+1}.$$

1. Sur la figure 1, on a représenté deux réalisations de (X_0, \dots, X_{100}) . L'une des deux correspond à $a = 0.99, b = 0.01$, l'autre à $a = 0.01, b = 0.99$. À votre avis, laquelle correspond au premier choix, et laquelle correspond au second choix ?

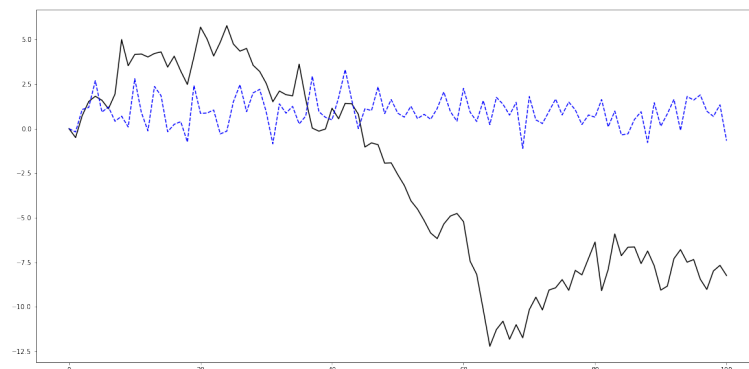


FIGURE 1 – Deux réalisations de (X_0, \dots, X_{100}) .

Correction — Dans le premier cas, a est proche de 1 et b proche de 0, donc la suite ressemble à une marche aléatoire qui fait des pas gaussiens. Ça doit beaucoup varier, on s'attend donc à ce que ce soit le tracé en trait plein.

Dans le second cas, les X_n sont « presque » des lois normales indépendantes centrées en $b \simeq 1$, ce qui correspond plutôt au tracé en pointillés.

2. On se demande si (X_n) converge vers une variable aléatoire X . Quel mode de convergence peut-on vraisemblablement exclure à partir de la figure 1 ?

Correction — Il semblerait que la convergence p.s. soit à exclure, au moins pour ces valeurs de a, b .

3. On veut écrire du code qui réalise la figure 1. On propose :

```
def X(n,a,b):
    x = 0
    for i in range(n-1):
        x = a*x + b + sts.norm.rvs()
    return x

listeX1 = [X(n,0.99,0.01) for n in range(101)]
listeX2 = [X(n,0.01,0.99) for n in range(101)]

plt.plot(listeX1)
plt.plot(listeX2)
plt.show()
```

Qu'en pensez-vous ?

Correction — Le problème de ce code est qu'il ne simule pas la suite (X_0, \dots, X_n) , mais des réalisations indépendantes $X_0, X_1^{(1)}, X_2^{(2)}$, etc. Chaque variable marginale a la bonne loi, mais la suite n'a pas la bonne loi. Ce type de code ne permet pas de tester la convergence p.s., car les variables sont indépendantes. Si on fait le dessin, la courbe en traits pleins va beaucoup plus « varier » que dans la figure 1...

4. Montrer que pour tout $n \geq 1$, $X_n \sim \mathcal{N}(\mu_n, \sigma_n^2)$, où μ_n, σ_n^2 sont à déterminer.

Correction — Par récurrence : $X_1 \sim \mathcal{N}(b, 1)$, et si $X_n \sim \mathcal{N}(\mu_n, \sigma_n^2)$, alors X_{n+1} étant indépendante de X_n (vu que X_n est mesurable par rapport à la tribu des (Z_1, \dots, Z_n) qui est indépendante de celle de Z_{n+1}), par les propriétés usuelles de la loi normale, $X_{n+1} \sim \mathcal{N}(a\mu_n + b, a^2\sigma_n^2 + 1)$. On trouve donc $\mu_{n+1} = a\mu_n + b$ et $\sigma_{n+1}^2 = a^2\sigma_n^2 + 1$. La résolution donne :

— si $a \neq 1$, $\mu_n = \frac{b(1-a^n)}{1-a}$; si $a = 1$, $\mu_n = nb$.

— si $|a| \neq 1$, $\sigma_n^2 = \frac{1-a^{2n}}{1-a^2}$; si $|a| = 1$, $\sigma_n^2 = n$.

Remarque : par la convention $\mathcal{N}(0, 0) = \delta_0$, ces résultats sont aussi vrais pour $n = 0$.

5. Pour quelles valeurs de a, b la suite (X_n) converge-t-elle en loi ? Dans ces cas-là, préciser sa limite.

Indication : on rappelle la fonction caractéristique d'une variable aléatoire $Z \sim \mathcal{N}(\mu, \sigma^2)$:

$$\phi_Z(t) = \mathbb{E}[e^{itZ}] = \exp\left(\mu it - \frac{\sigma^2 t^2}{2}\right).$$

On rappelle aussi le théorème de Lévy :

Une suite de v.a. réelles (Y_n) converge en loi vers Y ssi $\forall t \in \mathbb{R}, \phi_{Y_n}(t) \rightarrow \phi_Y(t)$.

Correction — Si $|a| < 1$, alors $\mu_n \rightarrow \frac{b}{1-a}$ et $\sigma_n^2 \rightarrow \frac{1}{1-a^2}$. Ainsi dans ce cas,

$$\phi_{X_n}(t) = \exp\left(\mu_n it - \frac{\sigma_n^2 t^2}{2}\right) \rightarrow \exp\left(\frac{b}{1-a} it - \frac{t^2}{2(1-a^2)}\right)$$

et par le théorème de Lévy, X_n converge en loi vers une $\mathcal{N}\left(\frac{b}{1-a}, \frac{1}{1-a^2}\right)$.

Si $|a| \geq 1$, alors on voit que $\sigma_n^2 \rightarrow \infty$, et on en déduit que

$$\phi_{X_n}(t) \rightarrow \mathbb{1}_{t=0}.$$

Or la fonction à droite n'est pas une fonction caractéristique (une fonction caractéristique est toujours continue, par le théorème de convergence dominée par exemple), donc par la réciproque du théorème de Lévy, la suite ne converge pas en loi dans ce cas.

6. Proposer une manière d'illustrer la convergence en loi précédente lorsqu'elle a lieu.

Coder votre méthode. On pourra utiliser librement la fonction `X` définie à la question 3.

Correction — On peut utiliser une fonction de répartition empirique avec n assez grand (ici $n = 100$ avec $N = 500$ réalisations), et la comparer à la fonction de répartition de la loi normale attendue :

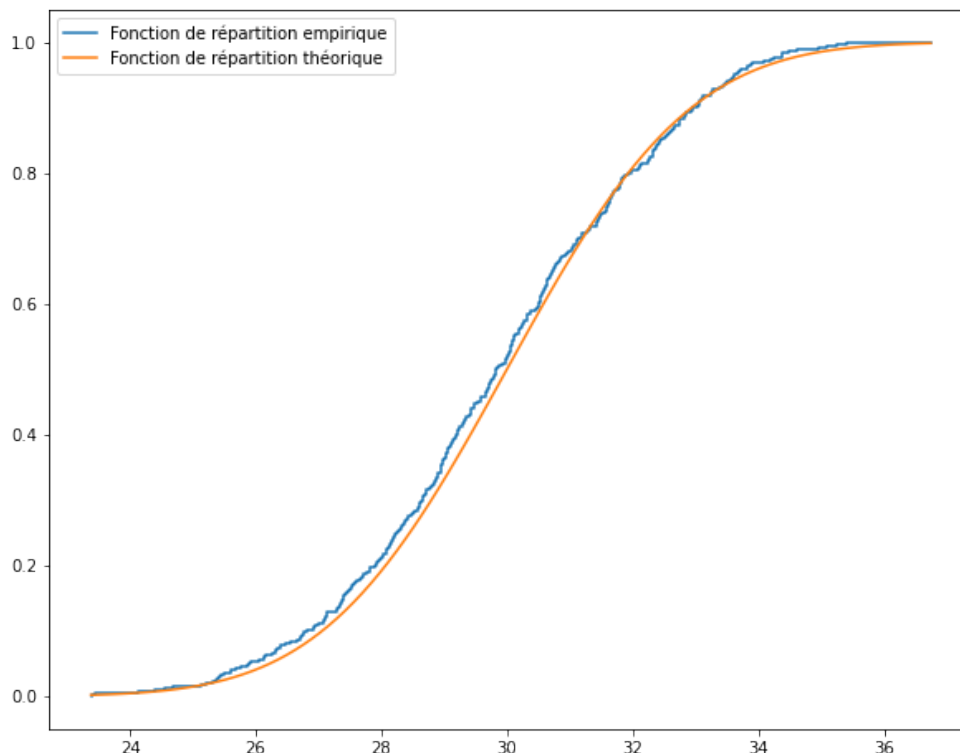
```
n=100
N=500
a=0.9
b=3

moy_lim = b / (1-a)
var_lim = 1 / (1-a**2)

donnees = [X(n,a,b) for _ in range(N)]
donnees_trie = sorted(donnees)
valeurs = np.linspace(0,1,N)
plt.step(donnees_trie,valeurs)

x = np.linspace( min(donnees), max(donnees), 100 )
y = sts.norm.cdf( x, loc=moy_lim, scale=np.sqrt(var_lim) )
# dans la ligne ci-dessus, le paramètre "scale" est l'écart-type, et non la variance !
plt.plot(x,y)

plt.show()
```



7. On se demande maintenant pour quelles valeurs de a, b la suite (X_n) converge en probabilité. Pourquoi est-ce difficile à tester avec l'ordinateur ?

Correction — Tester la convergence en proba requiert de connaître non seulement la loi X , mais aussi la loi jointe de (X_0, \dots, X_n, X) (ou en tout cas la loi de $X_n - X$). On sait bien simuler (X_0, \dots, X_n) et on connaît aussi la loi de X , mais on ne sait rien de leur loi jointe. Pour tester la convergence en loi, il faudrait déjà savoir comment les X_n et X sont couplées...

8. (*) Montrer que pour tout a, b , la suite (X_n) ne converge pas en probabilité.

Correction — Supposons par l'absurde que $X_n \rightarrow X$ en proba. Alors par les propriétés usuelles, $X_{n+1} - aX_n$ converge vers $(1-a)X$ en proba. Allons un peu plus loin : $(X_{n+3} - aX_{n+2}) - (X_{n+1} - aX_n)$ converge vers 0 en proba. Or cette variable est aussi égale à $Z_{n+3} - Z_{n+1}$, qui suit une loi $\mathcal{N}(0, 2)$ et converge donc en loi (vers... une $\mathcal{N}(0, 2)$!), elle ne peut donc pas converger en loi vers 0.

Exercice 3. — Méthode de Monte-Carlo et estimation de l'erreur

On souhaite estimer l'intégrale

$$I = \int_0^1 g(x) dx$$

où g est une fonction mesurable bornée connue. Pour cela on cherche une suite iid de variables aléatoires $(X_n)_{n \geq 1}$ à support borné, telle que $\mathbb{E}[X_1] = I$. On note $S_n = \sum_{k=1}^n X_k$.

1. Donner un exemple de variables (X_n) vérifiant ces hypothèses et facilement simulables.

Correction — Soient $(U_n)_{n \geq 1}$ des variables iid $\sim \mathcal{U}([0, 1])$, alors $X_n = g(U_n)$ conviennent (elles sont bornées par $\|g\|_\infty$ supposé fini, et leur espérance vaut I par transfert...).

2. On voudrait que $\frac{S_n}{n}$ soit une approximation de I à ϵ près. Rappeler brièvement pourquoi on doit prendre n d'ordre au moins ϵ^{-2} .

Correction — Par la loi forte des grands nombres, $\frac{S_n}{n}$ converge p.s. vers I , et par le théorème central limite,

$$\sqrt{n} \left| \frac{S_n}{n} - I \right|$$

converge en loi vers une $\mathcal{N}(0, \sigma^2)$ où $\sigma^2 = \text{Var}(X_1)$. Comme cette loi est intégrable, $\left| \frac{S_n}{n} - I \right|$ sera toujours d'ordre au moins $n^{-1/2}$. Pour que ce soit $\leq \epsilon$, il faut donc n d'ordre au moins ϵ^{-2} .

3. Plus précisément, on se donne un $\epsilon > 0$ et un $\alpha \in]0, 1[$. On cherche un n tel que

$$\mathbb{P} \left(\left| \frac{S_n}{n} - I \right| > \epsilon \right) \leq \alpha. \quad (*)$$

En utilisant Bienaymé-Tchebychev, donner un tel n en fonction de ϵ, α et de la fonction g .

Correction —

$$\begin{aligned} \mathbb{P} \left(\left| \frac{S_n}{n} - I \right| > \epsilon \right) &\leq \frac{1}{\epsilon^2} \text{Var} \left(\frac{S_n}{n} \right) \\ &\leq \frac{1}{\epsilon^2} \frac{n \text{Var}(X_1)}{n^2} \\ &\leq \frac{1}{n\epsilon^2} \left(\int_0^1 g^2(x) dx - I^2 \right). \end{aligned}$$

Ainsi il suffit de prendre $n = \lceil \frac{1}{\alpha\epsilon^2} \left(\int_0^1 g^2(x) dx - I^2 \right) \rceil$.

4. On admet le résultat suivant, appelé *lemme de Hoeffding* :

Soit Y une variable aléatoire à support dans $[a, b]$ et centrée ($\mathbb{E}[Y] = 0$), alors pour tout $s > 0$,

$$\mathbb{E}[e^{sY}] \leq \exp \left(\frac{s^2(b-a)^2}{8} \right).$$

- (a) En utilisant le lemme de Hoeffding, montrer que pour tout $s > 0$,

$$\mathbb{P}\left(\frac{S_n}{n} - I > \epsilon\right) \leq \exp(-ns\epsilon + Cns^2)$$

où C est une constante à déterminer qui ne dépend que de la fonction g .

Correction — Pour tout $s > 0$, en utilisant successivement l'inégalité de Markov, l'indépendance des X_n , et le fait que $X_1 - I$ est centrée et à valeurs dans $[-\|g\|_\infty - I, \|g\|_\infty - I]$:

$$\begin{aligned} \mathbb{P}\left(\frac{S_n}{n} - I > \epsilon\right) &= \mathbb{P}(\exp(s(S_n - nI)) > \exp(ns\epsilon)) \\ &\leq \exp(-ns\epsilon) \mathbb{E}[\exp(s(S_n - nI))] \\ &\leq \exp(-ns\epsilon) \mathbb{E}[\exp(s(X_1 - I))]^n \\ &\leq \exp(-ns\epsilon) \exp\left(n \frac{s^2 (2\|g\|_\infty)^2}{8}\right) \\ &\leq \exp(-ns\epsilon + Cns^2) \end{aligned}$$

où $C = \|g\|_\infty^2/2$.

- (b) En déduire que

$$\mathbb{P}\left(\frac{S_n}{n} - I > \epsilon\right) \leq \exp\left(\frac{-n\epsilon^2}{4C}\right).$$

Correction — On applique le résultat précédent en $s = \frac{\epsilon}{2C}$ (on peut le trouver comme le s qui minimise le polynôme de degré 2, $-ns\epsilon + Cns^2$).

- (c) Montrer alors que

$$\mathbb{P}\left(\left|\frac{S_n}{n} - I\right| > \epsilon\right) \leq 2 \exp\left(\frac{-n\epsilon^2}{4C}\right).$$

Correction —

$$\mathbb{P}\left(\left|\frac{S_n}{n} - I\right| > \epsilon\right) = \mathbb{P}\left(\frac{S_n}{n} - I > \epsilon\right) + \mathbb{P}\left(\frac{S_n}{n} - I < -\epsilon\right).$$

On a déjà traité le premier terme. Pour le second, il suffit d'appliquer le résultat du (b) aux variables $(-X_n)$.

- (d) En déduire une nouvelle valeur de n qui satisfait (\star) , et comparer avec le résultat de la question 2.

Correction — On en déduit que $n = \lceil \frac{4C}{\epsilon^2} \ln(2/\alpha) \rceil = \lceil \frac{2\|g\|_\infty^2}{\epsilon^2} \ln(2/\alpha) \rceil$ convient.

Comme dans le 2 ce résultat est d'ordre ϵ^{-2} , mais la différence se fait sur l'ordre en α . Dans le premier cas on avait du $1/\alpha$, alors qu'ici on a du $-\ln \alpha$, ce qui est plus petit quand α devient petit.